AI-Powered Detection of Synthetic Identities in Banking Using Behavioral Patterns

Aneri Patel, Evan Gregor, Tirth Ganatra

August 2025

1 Introduction

With the shift toward a digital-first economy, banking and financial services are becoming heavily dependent on online transactions, mobile platforms, and electronic verification systems. While digital banking has made financial services quicker and more convenient, it has also introduced new kinds of risks. One of the most concerning is synthetic identity fraud. Unlike any other identity theft, where a criminal uses the details of an actual person, synthetic identity fraud is carried out by mixing real and fake information like names, birth dates, or unique ID numbers to build a completely new, false identity. These fake identities often behave like genuine customers, sometimes even building a credit history before committing fraud, which makes them difficult to catch using older detection methods.

This type of fraud is now considered one of the fastest-growing financial crimes. For example, the U.S. Federal Reserve estimates that synthetic identity fraud causes over $6billion in losses each year. And because banking networks around the world are now so closely connected, the problem isn't just limited to developed nations countries with expanding digital systems face the same danger. The problem is that older fraud detection systems, which mostly rely on fixed rules or past patterns, often fall short when faced with these newer and more sophisticated scams.

Building a reliable fraud detection system for synthetic identities brings several benefits: it can reduce financial losses by catching suspicious behaviour early, strengthen confidence in digital banking, and support regulators by offering clear AI-based explanations for flagged cases.

In recent years, machine learning (ML) and deep learning (DL) have been increasingly applied to fight fraud. Methods like Autoencoders, Isolation Forests, Random Forests, and Support Vector Machines (SVMs) have shown strong results in detecting unusual patterns within transaction data. More advanced methods, such as graph-based models and recurrent neural networks (RNNs), go a step further by capturing relationships between users and tracking activity over time. To make these systems easier to trust, explainability tools like SHAP (SHapley Additive Explanations) are now being used to show clearly why an AI model flagged a case.

No single model is flawless. Supervised methods, for instance, depend on large labelled datasets—but such data is hard to find when it comes to synthetic fraud. Many models also fall behind as fraud tactics evolve. Some approaches also lack the transparency needed in tightly regulated sectors like banking. On top of that, behavioural patterns are often ignored, even though they can be some of the clearest signals for identifying synthetic users who are trying to pass themselves off as genuine customers.

This project introduces an AI-driven fraud detection system that centres on user behaviour—such as spending habits, balance movements, and unusual transaction timings. To highlight these behavioural signals, the system will use feature engineering and apply unsupervised models like Autoencoders and Isolation Forests to uncover suspicious activity. To make the results more trustworthy, SHAP will be used to explain why a user was flagged. A dashboard will also be developed to help fraud analysts quickly review and investigate suspicious cases. By combining precise detection with easy-to-understand explanations, the system takes a proactive and transparent approach to addressing synthetic identity fraud

Figure 1: Fraud detection

2 Literature Review

Sr. No.	Author Name	Year	Algorithm/Method Used	Dataset Used	Parameters Measured	Result
1	Anirban Majumder	2025	Supervised ML, NLP, Behavioral Analytics, CNN, RNN, Reinforcement Learning	Real-world case studies (PayPal, Google, IBM, Esure)	Accuracy, Fraud Detection Speed, False Positives, Real-time Adaptability	High accuracy with real-time fraud detection capabilities
2	Diego Vallarino	2025	Hybrid Mix-of-Experts: RNN + Transformer + Autoencoder	Synthetic credit card dataset (500K txns, 1.5% fraud)	Accuracy, Precision, Recall	98.7% Accuracy, 94.3% Precision, 91.5% Recall
3	R. Sinha et.al.	2025	GCN, Spectral Clustering, Anomaly Classification	Synthetic datasets based on FinCEN	Precision, Recall, Identity Consistency Score	92% precision in clustered identity fraud detection
4	P. Malik et.al.	2025	Behavioral Biometrics, Device Fingerprinting, Typing & Mouse Rhythm Detection	Simulated logs (1500 sessions, 80 synthetic IDs)	Keystroke Delay, Mouse Smoothness, Session Entropy	87% detection of synthetic identities
5	Ruhul Q. Majumder	2025	Comparative ML Review, Anomaly Detection, Risk Pattern Modeling	Real-time public fraud datasets (Eurostat, Kaggle)	Accuracy, AUC, Time to Flag Fraud	Effective on imbalanced and streaming data
6	Megha Adhikari et.al.	2024	AI-based Risk Profiling, ML, Behavioral Data Analytics, Transaction Monitoring	Banking and e-commerce data (unspecified)	Fraud Prevention Efficiency, Risk Prediction, Customer Protection, Regulatory Compliance	Enhanced fraud detection and data interpretation
7	K. Thakur et.al.	2024	Stream Classification, Isolation Forest, Online Logistic Regression	Streaming bank txn sim. (8,000 sequences)	True Positive Rate, Drift Resistance, Response Latency	90%+ detection, strong pattern-shift robustness
8	Wei Min et al.	2021	DeepBehaviorCluster, Bi-LSTM, Hybrid Feature Representation, pHDBSCAN, Skope-Rule	eBay transaction data (China & USA)	Time Cost, Precision, Recall, F-score, Return Rate	500x efficiency boost, F-score 23.52 with hybrid features

Sr. No.	Author Name	Year	Algorithm/Method Used	Dataset Used	Parameters Measured	Result
9	Giulia Moschini et al.	2021	ARIMA, Rolling Windows, Z-Score, K-means, LOF, Isolation Forest	Credit card data from NetGuardians SA	Precision, Recall, F1, ADF Test, Z-Score Threshold	Precision 34.29%–50%, adaptable to dynamic behavior
10	Mohammad M. R. Mashhadi	2019	Autoencoder, One-class SVM, Mahalanobis Distance	Real-world credit card data	AUC, Sensitivity, Specificity, RMSE	Autoencoder performed best; Mahalanobis useful without labels

Table 1: Literature Review

3 Proposed Methodology

3.1 Problem Statement and Objective

Synthetic identity fraud is a major challenge for the banking industry as it combines real and fabricated personal information to create a convincing but fraud customer profile. These scammers are able to bypass conventional verification systems, allowing them to open accounts, build transaction histories and execute large-scale financial crimes. Traditional fraud detection systems often fail to detect such scammers because of their legitimate activity patterns.

The objective of this study is to develop an AI-powered detection systems that learns behavioural patterns to accurately identify synthetic identities in banking. By analyzing the frequency of transactions, login times, device fingerprints ,behavioural patterns such as typing and mouse movements, the proposed model aims to detect these synthetic identities in real time while minimizing the false positives.

3.2 Research Design

This research makes use of quantitative, experimental design involving both supervised and unsupervised machine learning techniques. The Supervised models would be used to identify whether the identities are genuine or synthetic based on labelled datasets, whilst unsupervised clustering techniques will be used to understand the fraud patterns. This type of approach is influenced by R. Sinha and T. Morar (2025), who applied Graph Convolution Neural Networks for identity clustering, and Wei Min et. al.(2021), who demonstrated the efficiency of hybrid feature- based fraud detection.

This experimental nature of research allows testing in a systematic way using different algorithms to determine which model yields the highest accuracy, recall and adaptability for real world banking environments.

3.3 Population and Sample

The banking customers conducting online and offline transactions would be the population for this study. Due to confidentiality constraints, raw banking customer data would not be used by this research instead it would rely on anonymized public datasets and synthetically generated

identities. The sample will include diverse characteristics for robust model training following the approach of R. Sinha and T. Morar (2025)

The sample size would be determined on the basis of availability of public datasets and generatability of these realistic synthetic identities through simulation tools. Stratified sampling will be used to maintain equality of both genuine and fraudulent profiles

3.4 Data Collection Methods

Publicaly available fraud detection datasets would be used as the source of data (eg. Kaggle, IEEE, Data port) and will be augmented with synthetic identities generated using FinCEN guildelines. The required behavioural features will be extracted from transaction records, login logs, device history, geolocation data, and biometric activities such as keystroke patterns and mouse movements, as applied by P. Malim and R. Das (2025).

The dataset will capture both short-term and long-term behavioural patterns which would enable AI model to detect fraudulent activities that change over time. The use of synthetic behavioural logs ensures compliance with privacy regulations while providing realistic data for analysis.

3.5 Data Preprocessing

To ensure quality and consistency the dataset will go under preprocessing before training the model. Data cleaning will be done to remove the duplicates, correcting missing values and eliminating irrelevant attributes. Feature engineering will be used to generate new metrics such as transaction velocity, session entropy, and geolocation deviation.

Since class imbalance is a common issue in these types of datasets, techniques such as SMOTE (Synthetic Minority Oversampling Technique) or undersampling will be used. The categorical will be label encoded, and numerical features will be normalized to improve model compatibility and convergence speed.

3.6 AI Model Development

Multiple machine learning and deep learning models will be evaluated for detecting synthetic identities by this research. Supervised models such as Random Forest and XGBoost will be compared with deep learning models like Autoencoders and RNN + Transformer hybrids, by observing the high performance results obtained by Diego Vallarino (2025).

Unsupervised clustering methods such as HDBSCAN (Wei Min et. al., 2021), will be used to identify hidden groupings which may indicate synthetic profiles. Models will trained on 70% of the dataset and tested on 30% of the dataset and will be optimized through hyperparameter tuning.

3.7 Data Analysis Techniques

The first step would be Exploratory Data Analysis (EDA) to visualize trends , detect anomalies, and identify correlations among features. The behavioural indiactors which would contribute the most to fraud detection will be determined using feature importance analysis, similar to methods in Ruhul Q. Majumder (2025).

Performance evaluation will be done with metrics such as Precision, Recall , F1 – Score, and ROC-AUC. Following the recommendations of K. Thakur and L. Zhou (2024) detection latency and adaptability to behavioral drift will be measured. To evaluate stability across multiple sessions, the Identity Consistency Score (R. Sinha and T. Morar, 2025) will be applied.

3.8 Ethical Considerations

To ensure compliance with privacy regulation such as GDPR, all datasets will be anonymized to remove any personally identifiable information (PII) . Only synthetic identities and anonymized behavioral logs will be used in model training in order to avoid compromising real customer data.

The research will also ensure that the detection system does not bias towards demographic groups, aligning with responsible AI principles. The methodology follows the ethical framework of P. Malik and R. Das (2025), who demonstrated that behavioral biometrics can be analyzed without infringing on user privacy.

3.9 Expected Outcome

The proposed AI-powered fraud detection system is expected to achieve high accuracy and recall in detecting the synthetic identities in banking while also maintaining low false positive rates. The model should adapt to evolving fraud tactics and maintain stability over time by leveraging a hybrid of transactional and behavioral biometric features.

The research outcome will contribute to the development of real-time fraud detection frameworks in the banking industry, reducing financial losses and enhancing security. Additionally, the findings may be extended to related industries such as e-commerce and telecommunications, where synthetic identity fraud is also prevalent.

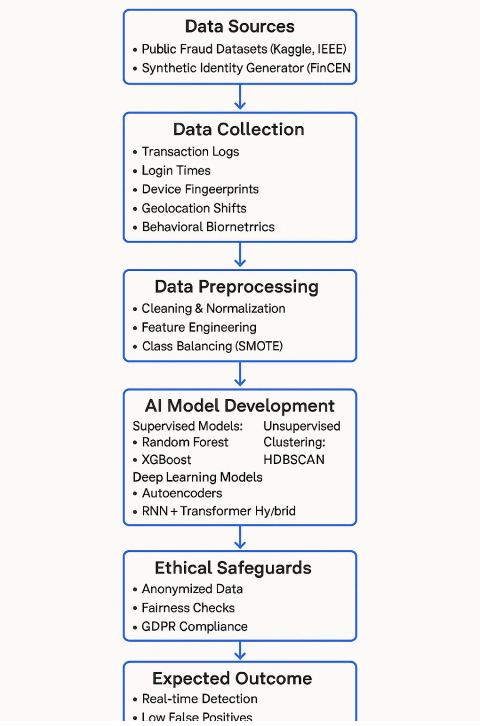

Figure 2: Fraud detection

References

[1] Majumder, A. (2025). Supervised & Unsupervised ML, NLP, Behavioral Analytics, CNN, RNN, Reinforcement Learning for Fraud Detection. *SSRG International Journal of Computer Science and Engineering, 12*(4), 17–22.

[2] Adhikari, M., Sharma, K., & Sharma, D. C. (2024). AI-based Risk Profiling and Transaction Monitoring Systems for Fraud Detection. *International Journal of Science and Research Archive.*

[3] Vallarino, D. (2025). Hybrid Mix-of-Experts (RNN + Transformer + Autoencoder) for Fraud Detection. *arXiv preprint arXiv:2504.03750v1.*

[4] Mashhadi, M. M. R. (2019). Unsupervised Fraud Detection Models: Autoencoder, One-class SVM, and Mahalanobis Distance. *International Journal of Advanced Computer Science and Applications, 10*(11).

[5] Min, W., Liang, W., Yin, H., Wang, Z., Li, M., & Lal, A. (2021). FinDeepBehaviorCluster: A Hybrid Feature Representation with GPU-powered HDBSCAN for Fraud Detection. *arXiv preprint arXiv:2101.04285v1.*

[6] Moschini, G., Houssou, R., Bovay, J., & Robert-Nicoud, S. (2021). Time-Series Fraud Detection using ARIMA, Rolling Windows, and Z-Score Methods. *Engineering Proceedings, 5*(56).

[7] Sinha, R., & Morar, T. (2025). Graph Convolutional Networks and Spectral Clustering for Synthetic Identity Fraud. *ResearchGate Preprint (DOI pending).*

[8] Malik, P., & Das, R. (2025). Behavioral Biometrics and Device Fingerprinting for Fraud Detection. *ResearchGate Preprint.*

[9] Thakur, K., & Zhou, L. (2024). Stream-based Classification and Isolation Forest for Fraud Detection. *ResearchGate Preprint.*

[10] Majumder, R. Q. (2025). Comparative Machine Learning Review and Anomaly Detection for Fraud. *SSRN Working Paper.*